Automatic Evaluation of Machine Translation Quality

Lluís Màrquez
TALP Research Center
Tecnhical University of Catalonia (UPC)

Invited talk at **Dialogue 2013**
Bekasovo Resort, Russia
May 30, 2013

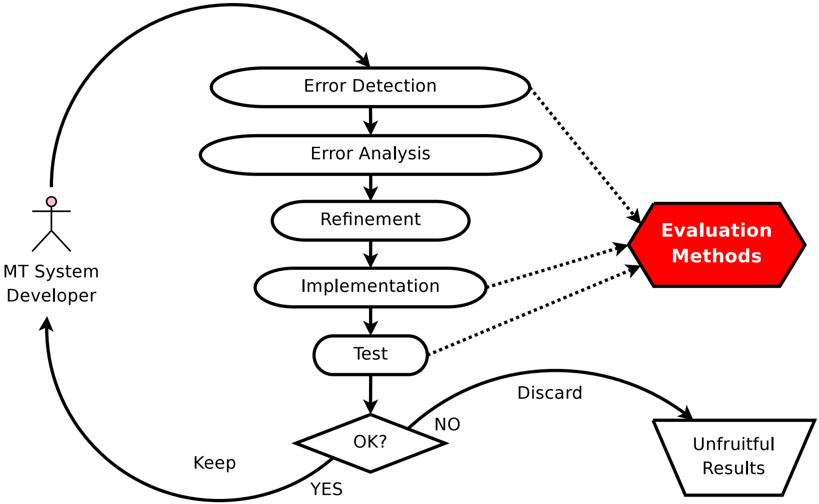
Joint work with:

Jesús Giménez, Lluís Formiga and Meritxell Gonzàlez

1. Automatic MT Evaluation
2. Linguistically-motivated Measures
3. Intelligent MT output and error analysis
4. Quality Estimation

Talk Overview

MT System Development Cycle



Difficulties of MT Evaluation

- Machine Translation is an *open* NLP task
  - ⇒ the *correct translation* is not unique
  - ⇒ the set of valid translations is not small
  - ⇒ translation correctness is not black and white

- Quality aspects are *heterogeneous*
  - ⇒ Adequacy (or Fidelity)
  - ⇒ Fluency (or Intelligibility)
  - ⇒ Post-editing effort (time, key strokes, ...)
  - ⇒ ...

- Manual vs. automatic evaluation

MT Automatic Evaluation

Setting:

⇒ Compute similarity between system's output and one or several reference translations

⇒ The similarity measure should be able to discriminate whether the two sentences convey the same meaning (semantic equivalence)

MT Automatic Evaluation

Setting:

⇒ Compute similarity between system's output and one or several reference translations

Challenge:

⇒ The similarity measure should be able to discriminate whether the two sentences convey the same meaning (semantic equivalence)

MT Automatic Evaluation

First Approaches:

$\Rightarrow$ Lexical similarity as a measure of quality

MT Automatic Evaluation

First Approaches:

$\Rightarrow$ Lexical similarity as a measure of quality

- **Edit Distance**
  WER, PER, TER
- **Precision**
  BLEU, NIST, WNM
- **Recall**
  ROUGE, CDER
- **Precision/Recall**
  GTM, METEOR, BLANC, SIA

MT Automatic Evaluation

First Approaches:

$\Rightarrow$ Lexical similarity as a measure of quality

- **Edit Distance**
  WER, PER, TER
- **Precision**
  BLEU, NIST, WNM
- **Recall**
  ROUGE, CDER
- **Precision/Recall**
  GTM, METEOR, BLANC, SIA

- **BLEU** has been widely accepted as a *'de facto'* standard

IBM BLEU metric

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu
IBM Research Division

"The main idea is to use a **weighted average of variable length phrase matches** against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family."

IBM BLEU metric

Conclusions of the paper (Papineni et al., 2001)

- BLEU correlates with human judgements
- It can distinguish among similar systems
- Need for multiple references or a big test with heterogeneous references
- More parametrisation in the future

Benefits of Automatic Evaluation

Compared to manual evaluation, automatic measures are:

1. Cheap (vs. costly)
2. Objective (vs. subjective)
3. Reusable (vs. not-reusable)

Automatic evaluation metrics have notably accelerated the development cycle of MT systems

1. Error analysis
2. System optimization
3. System comparison

Benefits of Automatic Evaluation

Compared to manual evaluation, automatic measures are:

1. Cheap (vs. costly)
2. Objective (vs. subjective)
3. Reusable (vs. not-reusable)

Automatic evaluation metrics have notably accelerated the development cycle of MT systems

1. Error analysis
2. System optimization
3. System comparison

Risks of Automatic Evaluation (compared to manual evaluation)

1. **System overtuning** → when system parameters are adjusted towards a given metric
2. **Blind system development** → when metrics are unable to capture actual system improvements
3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems

Risks of Automatic Evaluation (compared to manual evaluation)

1. **System overtuning** → when system parameters are adjusted towards a given metric
2. Blind system development → when metrics are unable to capture actual system improvements
3. Unfair system comparisons → when metrics are unable to reflect difference in quality between MT systems

Risks of Automatic Evaluation (compared to manual evaluation)

1. System overtuning → when system parameters are adjusted towards a given metric
2. **Blind system development** → when metrics are unable to capture actual system improvements
3. Unfair system comparisons → when metrics are unable to reflect difference in quality between MT systems

Risks of Automatic Evaluation (compared to manual evaluation)

1. System overtuning → when system parameters are adjusted towards a given metric
2. Blind system development → when metrics are unable to capture actual system improvements
3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems

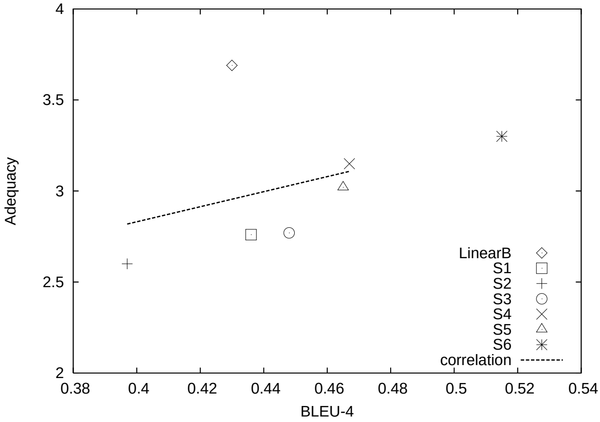
Problems of Lexical Similarity Measures

- Lexical similarity is nor a *sufficient* neither a *necessary* condition so that two sentences express the same meaning (Culy and Riehemann, 2003; Coughlin, 2003; Callison-Burch et al., 2006)
- The reliability of lexical metrics depends very strongly on the heterogeneity/representativity of reference translations
- Lexical metrics have problems distinguishing MT output from fully fluent and adequate translations obtained from them through professional postediting (Denkowski and Lavie, 2012)

Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise

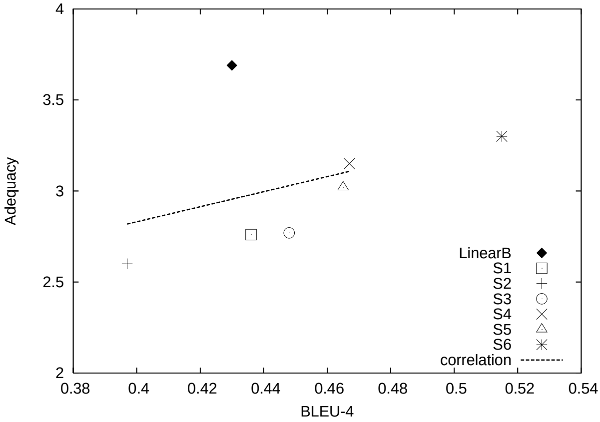
(Callison-Burch et al., 2006; Koehn and Monz, 2006)



Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise

(Callison-Burch et al., 2006; Koehn and Monz, 2006)



Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)

⇒ *n*-gram based metrics favor MT systems which closely replicate the lexical realization of the references

⇒ Test sets tend to be similar (domain, register, sublanguage) to training materials

⇒ Statistical MT systems heavily rely on the training data

⇒ Statistical MT systems tend to share the reference sublanguage and be favored by *n*-gram based measures

Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise

(Callison-Burch et al., 2006; Koehn and Monz, 2006)

- ⇒ *n*-gram based metrics favor MT systems which closely replicate the lexical realization of the references
- ⇒ Test sets tend to be similar (domain, register, sublanguage) to training materials
- ⇒ Statistical MT systems heavily rely on the training data
- ⇒ Statistical MT systems tend to share the reference sublanguage and be favored by *n*-gram based measures

Talk Overview

Can we do better?

1. Compare to a very large set of references
   - HyTER (Dreyer and Marcu, 2012)
     - ⇒ Construct for every test case a compact network encoding an exponentially large number of meaning equivalent reference translations
     - ⇒ Compute a TER-based similarity over the whole set of translation equivalents
     - ⇒ HyTER correlates much better with human assessments
     - ⇒ But the cost of generating the graphs is very high

Can we do better?

1. Compare to a very large set of references

- HyTER (Dreyer and Marcu, 2012)
  - ⇒ Construct for every test case a compact network encoding an exponentially large number of meaning equivalent reference translations
  - ⇒ Compute a TER-based similarity over the whole set of translation equivalents
  - ⇒ HyTER correlates much better with human assessments
  - ⇒ But the cost of generating the graphs is very high

Can we do better?

2. **Generalize over lexical matching**

- Lexical variants
  - ⇒ Morphological information (i.e., stemming) ROUGE and METEOR
  - ⇒ Synonymy lookup: METEOR (based on WordNet)
- Paraphrasing support:
  - ⇒ (Zhou et al., 2006; Kauchak and Barzilay, 2006; Owczarzak et al., 2006)
  - ⇒ Recent versions of METEOR, TER

Similarity Measures Based on Linguistic Features

3. More linguistically-motivated measures

- Features capturing syntactic and semantic information
- Shallow parsing, constituency and dependency parsing, named entities, semantic roles, textual entailment, discourse representation
- Very extense bibliography in the last years
  Check (Giménez and Màrquez 2010) for a survey

Some Examples of Linguistically Motivated Measures

- Expected Dependency Pair Match
  (Kahn, Snover and Ostendorf, 2009)
  - ⇒ dependency parsing (PCFG + head-finding rules)
  - ⇒ precision and recall scores of various tree decompositions
  - ⇒ +synonymy +paraphrasing

- MaxSim (Chen and Ng; 2008)
  - ⇒ a general framework for arbitrary similarity functions
  - ⇒ dependency relations, lemma, parts of speech, synonymy
  - ⇒ bipartite graph to obtain an optimal matching between items

- RTE (Padó, Galley, Jurafsky and Manning, 2009)
  - ⇒ semantic equivalence based on textual entailment features
  - ⇒ alignment, semantic compatibility, insertion/deletion, preservation of reference and structural alignment

Some Examples of Linguistically Motivated Measures

- Expected Dependency Pair Match
  (Kahn, Snover and Ostendorf, 2009)
  - $\Rightarrow$ dependency parsing (PCFG + head-finding rules)
  - $\Rightarrow$ precision and recall scores of various tree decompositions
  - $\Rightarrow$ +synonymy +paraphrasing

- MaxSim (Chen and Ng; 2008)
  - $\Rightarrow$ a general framework for arbitrary similarity functions
  - $\Rightarrow$ dependency relations, lemma, parts of speech, synonymy
  - $\Rightarrow$ bipartite graph to obtain an optimal matching between items

- RTE (Padó, Galley, Jurafsky and Manning, 2009)
  - $\Rightarrow$ semantic equivalence based on textual entailment features
  - $\Rightarrow$ alignment, semantic compatibility, insertion/deletion, preservation of reference and structural alignment

Our Approach

(Giménez & Màrquez, 2010)

Work at UPC with Jesús Giménez

Rather than comparing sentences at lexical level:

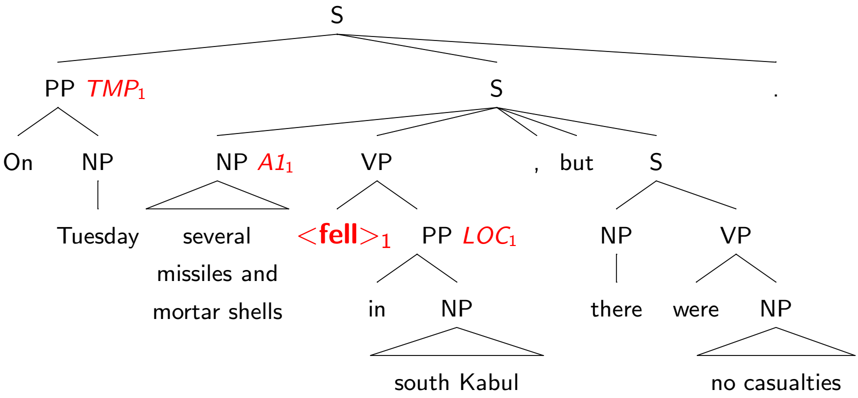
Compare the linguistic structures and the words within them

Our Approach

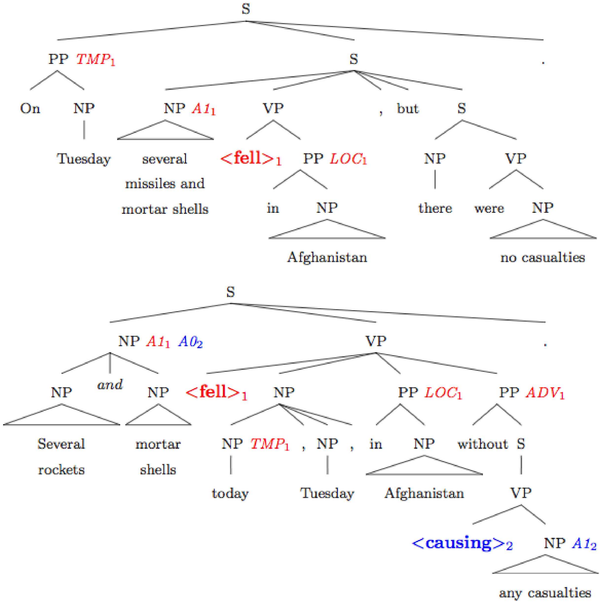
(Giménez & Màrquez, 2010)

Automatic Translation	On Tuesday several missiles and mortar shells fell in south Kabul , but there were no casualties .
Reference Translation	Several rockets and mortar shells fell today , Tuesday , in south Kabul without causing any casualties .

Our Approach (Giménez & Màrquez, 2010)



Our Approach (Giménez & Màrquez, 2010)



Measuring Structural Similarity

- OVERLAP: generic similarity measure among Linguistic Elements. Inspired by the Jaccard similarity coefficient

- Linguistic element (LE) = abstract reference to any possible type of linguistic unit, structure, or relationship among them
  - ⇒ For instance: POS tags, word lemmas, NPs, syntactic phrases
  - ⇒ A sentence can be seen as a bag (or a sequence) of LEs of a certain type
  - ⇒ LEs may embed

Measuring Structural Similarity

- OVERLAP: generic similarity measure among Linguistic Elements. Inspired by the Jaccard similarity coefficient
- Linguistic element (LE) = abstract reference to any possible type of linguistic unit, structure, or relationship among them
  - ⇒ For instance: POS tags, word lemmas, NPs, syntactic phrases
  - ⇒ A sentence can be seen as a bag (or a sequence) of LEs of a certain type
  - ⇒ LEs may embed

Overlap among Linguistic Elements

$$\mathrm{O}(t) = \frac{\displaystyle\sum_{i \in (\mathrm{items}_t(\mathrm{hyp})\ \cap\ \mathrm{items}_t(\mathrm{ref}))} \mathrm{count}_{\mathrm{hyp}}(i,t)}{\displaystyle\sum_{i \in (\mathrm{items}_t(\mathrm{hyp})\ \cup\ \mathrm{items}_t(\mathrm{ref}))} \max(\mathrm{count}_{\mathrm{hyp}}(i,t), \mathrm{count}_{\mathrm{ref}}(i,t))}$$

t is the LE type
'hyp': hypothesized translation
'ref': reference translation
$\mathrm{items}_t(s)$: set of items occurring inside LEs of type t
$\mathrm{count}_s(i,t)$: occurrences of item i in s inside a LE of type t

Overlap among Linguistic Elements

Coarser variant: micro-averaged overlap over all types

$$O(\star) = \frac{\sum_{t \in T} \sum_{i \in (\text{items}_t(\text{hyp}) \,\cap\, \text{items}_t(\text{ref}))} \text{count}_{\text{hyp}}(i, t)}{\sum_{t \in T} \sum_{i \in (\text{items}_t(\text{hyp}) \,\cup\, \text{items}_t(\text{ref}))} \max(\text{count}_{\text{hyp}}(i, t), \text{count}_{\text{ref}}(i, t))}$$

T: set of all LE types associated to the given LE class

Overlap/Matching among Linguistic Elements

- Matching is a similar but more strict variant
  - ⇒ All items inside an element are considered the same unit
  - ⇒ Computes the proportion of fully translated LEs, according to their types
- Other possible extensions:
  - ⇒ *n*-gram matching within LEs
  - ⇒ Synonymy lookup

Overlap/Matching among Linguistic Elements

- Matching is a similar but more strict variant
  - ⇒ All items inside an element are considered the same unit
  - ⇒ Computes the proportion of fully translated LEs, according to their types
- Other possible extensions:
  - ⇒ *n*-gram matching within LEs
  - ⇒ Synonymy lookup

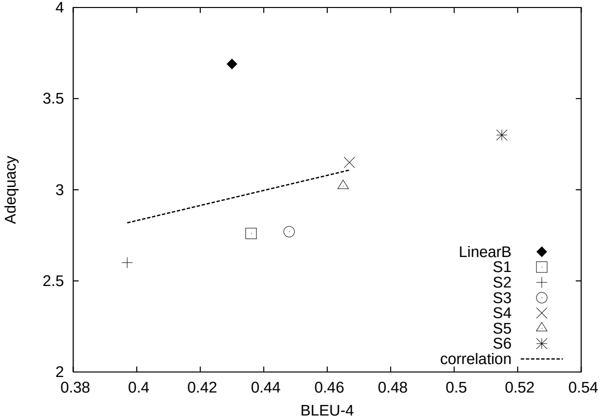
Overlap/Matching among Linguistic Elements

- Overlap and Matching have been instantiated over different linguistic level elements (for English)
  - ⇒ Words, lemmas, POS
  - ⇒ Shallow, dependency and constituency parsing
  - ⇒ Named entities and semantic roles
  - ⇒ Discourse representation (logical forms)

Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

(Callison-Burch et al., 2006; Koehn and Monz, 2006)



Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

Level	Metric	ρ_{all}	ρ_{SMT}
Lexical	BLEU	0.06	0.83
	METEOR	0.05	**0.90**
Syntactic	Parts-of-speech	0.42	0.89
	Dependencies (HWC)	**0.88**	0.86
	Constituents (STM)	0.74	**0.95**
Semantic	Semantic Roles	0.72	**0.96**
	Discourse Repr.	0.92	0.92
	Discourse Repr. (PoS)	**0.97**	0.90

Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

Level	Metric	ρ_{all}	ρ_{SMT}
Lexical	BLEU	0.06	0.83
	METEOR	0.05	**0.90**
Syntactic	Parts-of-speech	0.42	0.89
	Dependencies (HWC)	**0.88**	0.86
	Constituents (STM)	0.74	**0.95**
Semantic	Semantic Roles	0.72	**0.96**
	Discourse Repr.	0.92	0.92
	Discourse Repr. (PoS)	**0.97**	0.90

Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

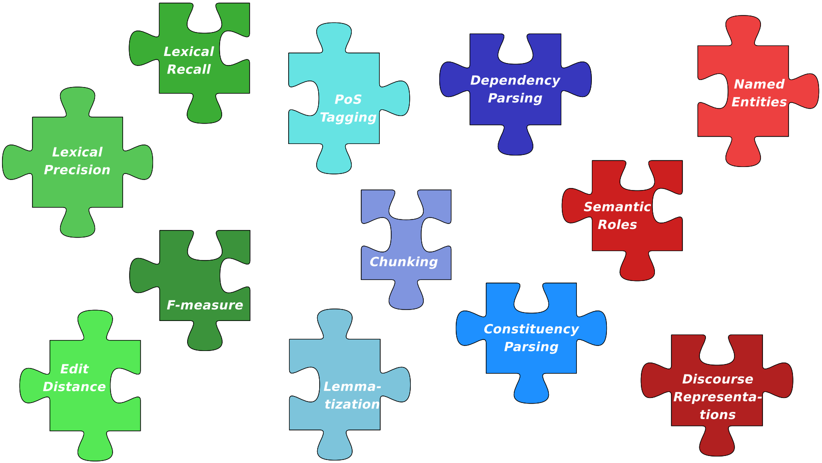
Level	Metric	ρ_{all}	ρ_{SMT}
Lexical	BLEU	0.06	0.83
	METEOR	0.05	**0.90**
Syntactic	Parts-of-speech	0.42	0.89
	Dependencies (HWC)	**0.88**	0.86
	Constituents (STM)	0.74	**0.95**
Semantic	Semantic Roles	0.72	**0.96**
	Discourse Repr.	0.92	0.92
	Discourse Repr. (PoS)	**0.97**	0.90

Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

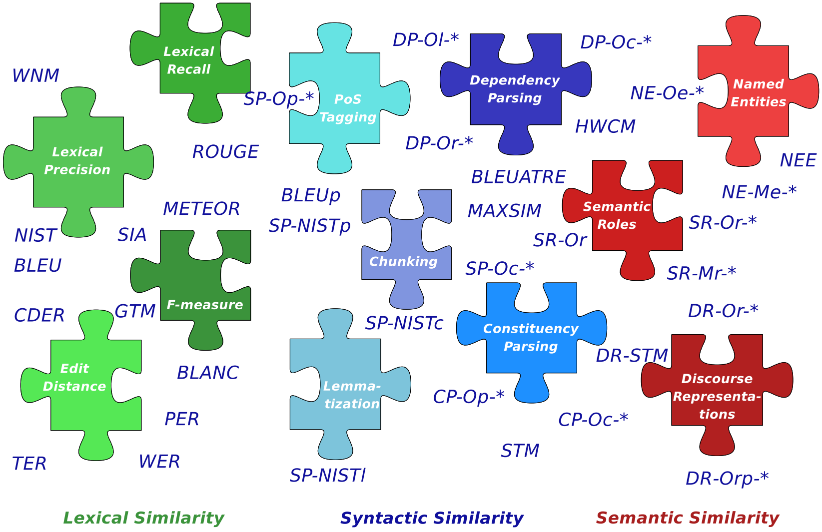
Level	Metric	ρ_{all}	ρ_{SMT}
Lexical	BLEU	0.06	0.83
	METEOR	0.05	**0.90**
Syntactic	Parts-of-speech	0.42	0.89
	Dependencies (HWC)	**0.88**	0.86
	Constituents (STM)	0.74	**0.95**
Semantic	Semantic Roles	0.72	**0.96**
	Discourse Repr.	0.92	0.92
	Discourse Repr. (PoS)	**0.97**	0.90

Towards Heterogeneous Automatic MT Evaluation



Lexical Similarity ***Syntactic Similarity*** ***Semantic Similarity***

Towards Heterogeneous Automatic MT Evaluation



Combined Evaluation Measures

- Different measures capture different aspects of similarity
  Suitable for combination
- Extense bibliography on learning to combine evaluation measures. Check (Giménez and Màrquez 2010) for a survey

The Most Simple Approach: ULC

- Uniformly averaged linear combination of measures (ULC):

$$\mathrm{ULC}_M(hyp, ref) = \frac{1}{|M|} \sum_{m \in M} m(hyp, ref)$$

- Simple hill climbing approach to find the best subset of measures M on a development corpus

$M = \{$ '$ROUGE_W$', 'METEOR', '$DP\text{-}HWC_r$', '$DP\text{-}O_c(\star)$', '$DP\text{-}O_l(\star)$', '$DP\text{-}O_r(\star)$', '$CP\text{-}STM_4$', '$SR\text{-}O_r(\star)$', '$SR\text{-}O_{rv}$', '$DR\text{-}O_{rp}(\star)$' $\}$

The Most Simple Approach: ULC

- Uniformly averaged linear combination of measures (ULC):

$$\mathrm{ULC}_M(hyp, ref) = \frac{1}{|M|} \sum_{m \in M} m(hyp, ref)$$

- Simple hill climbing approach to find the best subset of measures M on a development corpus

$M = \{$ '$ROUGE_W$', 'METEOR', '$DP\text{-}HWC_r$', '$DP\text{-}O_c(\star)$', '$DP\text{-}O_l(\star)$', '$DP\text{-}O_r(\star)$', '$CP\text{-}STM_4$', '$SR\text{-}O_r(\star)$', '$SR\text{-}O_{rv}$', '$DR\text{-}O_{rp}(\star)$' $\}$

Evaluation of ULC

WMT 2008 meta-evaluation results (into-English)

Measure	ρ_{sys}	consistency$_{snt}$
ULC	**0.83**	**0.56**
DP-O_r($\star$)	**0.83**	0.51
DR-O_r($\star$)	0.80	0.50
METEOR$_{ranking}$	0.78	0.51
SR-O_r($\star$)	0.77	0.50
METEOR$_{baseline}$	0.75	0.51
PoS-BLEU	0.75	0.44
PoS-4gram-F	0.74	0.50
BLEU	0.52	—
BLEU$_{stem+wnsyn}$	0.50	0.51
...		

Evaluation of ULC

WMT 2009 meta-evaluation results (into-English)

Measure	ρ_{sys}	consistency$_{snt}$
ULC	**0.83**	**0.54**
maxsim	0.80	0.52
rte(absolute)	0.79	0.53
meteor-rank	0.75	0.49
rte(pairwise)	0.75	0.51
terp	-0.72	0.50
meteor-0.6	0.72	0.49
meteor-0.7	0.66	0.49
bleu-ter/2	0.58	—
nist	0.56	—
wpF	0.56	0.52
ter	-0.54	0.45
...		

Portability Across Corpora

NIST 2004/2005 MT Evaluation Campaigns

	AE_{2004}	**CE_{2004}**	**AE_{2005}**	**CE_{2005}**
#references	5	5	5	4
$\#outputs_{ass.}$	5/5	10/10	6/7	5/10
$\#sentences_{ass.}$	347/1,353	447/1,788	266/1,056	272/1,082
Avg. Adequacy	2.81/5	2.60/5	3.00/5	2.58/5
Avg. Fluency	2.56/5	2.41/5	2.70/5	2.47/5

Portability Across Corpora

Meta-evaluation of ULC across test beds
(Pearson Correlation)

	AE_{04}	CE_{04}	AE_{05}	CE_{05}
ULC ($_{AE04}$)	0.6392	0.6294	0.5327	0.5695
ULC ($_{CE04}$)	0.6306	0.6333	0.5115	0.5692
ULC ($_{AE05}$)	0.6175	0.6029	0.5450	0.5706
ULC ($_{CE05}$)	0.6218	0.6208	0.5270	0.6047
Max Indiv.	0.5877	0.5955	0.4960	0.5348

Linguistic Measures at International Campaigns

- Many MT evaluation campaigns have been conducted in the last years under NIST, WMT and IWSLT events
- Controversial results at NIST Metrics MATR08/09 Challenges, with bad results in general for linguistic–based evaluation measures
- Finding a practical robust automatic evaluation metric, which correlates well with human assessments is still an open problem

Summary

1. Evaluation methods play a crucial role
2. Measuring overall translation quality is hard
   - ⇒ Quality aspects are heterogeneous and diverse
3. What can we do?
   - ⇒ Advance towards heterogeneous evaluation methods
   - ⇒ Metricwise system development

     Always meta-evaluate
     (make sure your metric fits your purpose)
   - ⇒ Resort to manual evaluation

     Always conduct manual evaluations
     (contrast your automatic evaluations)
     Always do error analysis (semi-automatic)

Talk Overview

MT output and error analysis

ASIYA: An Open Toolkit for Automatic MT Evaluation

⇒ Integrates all the evaluation measures from (Giménez and Màrquez, 2010)

⇒ Goal: to facilitate a practical analysis of large and complex test suites, along several dimensions

- ▷ System evaluation and comparison with a rich family of metrics
- ▷ Error analysis
- ▷ Meta-evaluation of evaluation metrics

⇒ Useful for MT system and evaluation metric developers

⇒ Available and downloadable from: http://www.lsi.upc.es/~nlp/Asiya/

MT output and error analysis

Recent developments

⇒ ASIYA *in the cloud* (Gonzàlez et al., 2012;2013)

1. ASIYA Web Service
2. ASIYA Online Interface
3. ASIYA *t*SEARCH module

⇒ Demo video at the same ASIYA website

Talk Overview

Translation Quality Estimation

Quality Estimation (QE)

⇒ Estimate translation quality without reference translations

⇒ Information available

- ▷ Source sentence, candidate translation(s), and some MT system information

⇒ Application scenarios

- ▷ Informing MT end-users about estimated translation quality
- ▷ Quality-oriented filtering of translated texts
  - ⇒ identify translations requiring manual post-edition
  - ⇒ identify useful post-editions from users
- ▷ Ranking of several translation alternatives
  - ⇒ system selection, parameter optimization

Translation Quality Estimation

Quality Estimation (QE)

⇒ Estimate translation quality without reference translations

⇒ Information available

- ▷ Source sentence, candidate translation(s), and some MT system information

⇒ Application scenarios

- ▷ Informing MT end-users about estimated translation quality
- ▷ Quality-oriented filtering of translated texts
  - ⇒ identify translations requiring manual post-edition
  - ⇒ identify useful post-editions from users
- ▷ Ranking of several translation alternatives
  - ⇒ system selection, parameter optimization

Translation Quality Estimation

QE approaches

⇒ Scoring task to predict the absolute quality of the automatic translation of an input text

- ▷ Usually implemented as a regression function
- ▷ Also as a direct ranking between translation alternatives
- ▷ Supervised learning from a training set with human assessments

Translation Quality Estimation

Relevant work

⇒ Johns Hopkins University Summer Workshop, 2003
"Confidence Estimation for Machine Translation"
(Blatz et al., 2003)

⇒ Recent work:
(Specia et al., 2009;2010), (Soricut and Echihabi, 2010), (Giménez and Specia 2010), (Pighin et al., 2011), (Avramidis, 2012), etc.

⇒ WMT 2012 shared task on Quality Estimation
(Callison-Burch et al., 2012) (2nd edition at WMT 2013)

Quality Estimation

Features to train the QE measures

- System–dependent
- System–independent

Quality Estimation

Features to train the QE measures

- System–dependent
  - ⇒ internal system probabilities/scores
  - ⇒ features over *n*-best translation hypotheses
    - ▷ language modeling
    - ▷ hypothesis rank
    - ▷ score ratio
    - ▷ average hypothesis length
    - ▷ length ratio
    - ▷ center hypothesis
- System–independent

Quality Estimation

Features to train the QE measures

- System–dependent
- System–independent
  - $\Rightarrow$ Source (translation *difficulty*)
    - ▷ sentence length
    - ▷ ambiguity $\longrightarrow$ dictionary/alignment/WordNet-based (number of candidate translations per word or phrase)
  - $\Rightarrow$ Target (translation *fluency*)
    - ▷ sentence length
    - ▷ language modeling
  - $\Rightarrow$ Source-Target (translation *adequacy*)
    - ▷ length ratio
    - ▷ punctuation issues
    - ▷ candidate matching $\longrightarrow$ dictionary-/alignment-based

Translation Quality Estimation

QE challenges

⇒ QE is as difficult as MT itself!

⇒ Real adequacy–based QE measures are difficult to apply

- ▷ Training sets are small
- ▷ Involving sophisticated linguistic knowledge easily leads to severe data sparseness

The FAUST Project (2010-2013)

- Feedback Analysis for User Adaptive Statistical Translation
- FP7-ICT-2009-4 (Language-based interaction)
- http://divf.eng.cam.ac.uk/faust

Goal Develop interactive machine translation systems which adapt rapidly and intelligently to user feedback

- Challenges in FAUST: *real life MT*
  - ⇒ Open general translation
  - ⇒ Casual users (feedback is unreliable)
  - ⇒ Non-standard and noisy translation texts
  - ⇒ Rapid integration of feedback is required

The FAUST Project (2010-2013)

- Feedback Analysis for User Adaptive Statistical Translation
- FP7-ICT-2009-4 (Language-based interaction)
- http://divf.eng.cam.ac.uk/faust

Goal Develop interactive machine translation systems which adapt rapidly and intelligently to user feedback

- Challenges in FAUST: *real life MT*
  - ⇒ Open general translation
  - ⇒ Casual users (feedback is unreliable)
  - ⇒ Non-standard and noisy translation texts
  - ⇒ Rapid integration of feedback is required

Learning Quality Estimation measures (FAUST)

Task Training a combination of simple QE features to produce better predictors of translation quality on FAUST data

- Setting
  - ⇒ We used human feedback in the form of translation quality pairwise rankings. FAUST benchmark corpus: ∼1,900 input segments (en-es), translated by 5 MT systems
  - ⇒ Use of several feature families. Some novel
  - ⇒ Regression vs. ranking SVM learning
  - ⇒ Evaluation in terms of:
    - ▷ Correlation of the predicted rankings with the gold standard
    - ▷ Selection of the best translation (system combination)

Learning Quality Estimation measures (FAUST)

We considered features from 4 different families

1. Specia Baseline (17) (Specia et al., 2010)
   - token counts and their ratio, LM probabilities, *n*-grams filtered by quartiles, punctuation marks and fertility ratios
2. Asiya QE features (26) (Gonzàlez et al., 2012)
   - bilingual dictionary ambiguity and overlap; overlap ratios on chunks, named-entities and PoS; source and candidate language model perplexities and inverse perplexities over lexical forms, chunks and PoS and out-of-vocabulary word indicators
3. Features based on adapted Language Models (2)
   - Words and POS tags. Interpolation weights were computed as to minimize the perplexity according to the Spanish FAUST development set

Learning Quality Estimation measures (FAUST)

We considered features from 4 different families

4. Pseudo-reference based features (Soricut and Echihabi, 2010)
   - Idea: automatically produced translations by other systems are taken as references
   - Rationale: if system X produced a translation A and system Y produced a translation B starting from the same input, and A and B are similar and X and Y are different systems, then A is probably a good translation
   - Calculated with BLEU, NIST, METEOR, etc. (5) but also with the linguistic-based metrics from ASIYA (23)

Learning Quality Estimation measures (FAUST)

- Main Results on FAUST test data
  - ⇒ It is possible to learn reasonably good QE models from the FAUST annotated corpus, exhibiting fair correlation with the gold-standard rankings
  - ⇒ For the system selection task, pairwise ranking yields better results than regression
  - ⇒ Results are clearly over the baselines. They are also slightly over the system-informed Oracle-D(ominant)
  - ⇒ All proposed extensions of the basic feature set were useful to boost the quality of the QE modelssystem selection task

Learning Quality Estimation measures (FAUST)

- Quality of the predicted rankings
  - ⇒ Spearman correlation (ρ): 33.86 – **38.43**
  - ⇒ Kendall correlation (τ): 29.67 – **33.02**
  - ⇒ Accuracy of pairwise rankings: 44.67 – 58.11
  - ⇒ Accuracy at predicting best translation: 39.44 – **51.11**
- Results on the system selection task

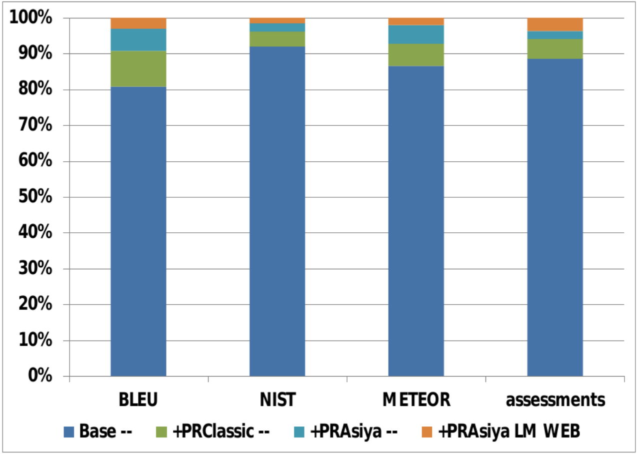
Learning Quality Estimation measures (FAUST)

- Quality of the predicted rankings
  - ⇒ Spearman correlation (ρ): 33.86 – **38.43**
  - ⇒ Kendall correlation (τ): 29.67 – **33.02**
  - ⇒ Accuracy of pairwise rankings: 44.67 – 58.11
  - ⇒ Accuracy at predicting best translation: 39.44 – **51.11**

- Results on the system selection task

	Baseline	**Ranker**	**OracleD**	**OracleB**
BLEU	33.64	**38.28**	37.57	44.91
METEOR	48.34	**54.19**	54.09	58.15
NIST	33.64	**38.28**	37.57	44.91

Learning Quality Estimation measures (FAUST)



Contribution of every family of features

Thank you!

Automatic Evaluation of Machine Translation Quality

Lluís Màrquez
TALP Research Center
Tecnhical University of Catalonia (UPC)

Invited talk at **Dialogue 2013**
Bekasovo Resort, Russia
May 30, 2013